TC3006C – Inteligencia artificial avanzada para la ciencia de datos I (Gpo 102)

Módulo 2 – Aprendizaje máquina (automático)

Análisis y reporte de desempeño del modelo

Profesor: César Javier Guerra Páramo

Edgar Castillo Ramírez A00827826

Monterrey, Nuevo León, 11 de septiembre de 2022.

I.- Implementación elegida

Para la resolución de esta actividad, se analizaron los modelos que se han visto en clase. Se seleccionó la implementación de un modelo de Árboles de Decisión Regresión, en la que se utilizaron librerías (Sklearn) para generarlo.

El conjunto de datos utilizado es describe las alturas y pesos de hombres y mujeres, es un conjunto que se nos dio en la clase de estadística. Cuenta con un total de 220 datos.

H_estat	H_peso	M_estat	M_peso
1.61	72.21	1.53	50.07
1.61	65.71	1.6	59.78
1.7	75.08	1.54	50.66
1.65	68.55	1.58	56.96
1.72	70.77	1.61	51.03
1.63	77.18	1.57	64.27
1.76	81.21	1.61	68.62
1.67	75.71	1.52	54.53
1.67	76.57	1.62	66.96
1.65	68.78	1.63	66.94
1.63	65.13	1.55	59.84
1.7	77.53	1.6	55.46
1.69	70.91	1.51	57.54
1.59	71.77	1.59	50.05
1.71	80.98	1.53	50.25
1.66	74.11	1.67	64.36
1.65	72.45	1.56	53.79

En la primera parte del código, se muestra el dataset, se imprimen algunos valores para observar cómo es la tabla y se revisa que los datos estén limpios. Es decir, se revisan valores faltantes y duplicados.

```
---------------VISUALIZACIÓN DE DATOS---------------

   H_estat  H_peso  M_estat  M_peso
0     1.61   72.21     1.53   50.07
1     1.61   65.71     1.60   59.78
2     1.70   75.08     1.54   50.66
3     1.65   68.55     1.58   56.96
4     1.72   70.77     1.61   51.03

---------------VALORES FALTANTES---------------
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 220 entries, 0 to 219
Data columns (total 4 columns):
 #   Column   Non-Null Count  Dtype
---  ------   --------------  -----
 0   H_estat  220 non-null    float64
 1   H_peso   220 non-null    float64
 2   M_estat  220 non-null    float64
 3   M_peso   220 non-null    float64
dtypes: float64(4)
memory usage: 7.0 KB
None

---------------VALORES DUPLICADOS---------------
Suma de duplicados: 0
```

Como se puede observar, el conjunto de datos no presentó valores faltantes ni duplicados, por lo que en ese aspecto no hay problemas.

Ahora bien, lo que se busca al generar el modelo es realizar predicciones, por lo que la pregunta de interés es: **¿Influye la estatura en el peso de un hombre?**

II.- Análisis del desempeño del modelo en un set de datos

Separación y evaluación del modelo con un conjunto de prueba y un conjunto de validación

Para esta parte se utilizó la librería. Si bien no contaba con un par de datasets de prueba y validación, es posible generarlos a partir de los 220 registros. Por esta razón, un 20% de los datos son de prueba y el resto de entrenamiento. El código brinda una impresión de las dimensiones de los datos una vez se realiza la separación de los datos:

```
---------------DIMENSIONES DE DATOS DE ENTRENO Y PRUEBA---------------
De 220 datos dentro de la tabla:
x_train: (176, 1)
x_test: (44, 1)
y_train: (176,)
y_test: (44,)
```

Diagnóstico y explicación del grado de bias o sesgo

Para este apartado se utilizó el Mean Bias Error o el error de sesgo promedio. Con los datos encontrados. Si recordamos lo que es un Árbol de decisión, es esperado que no exista una gran cantidad de sesgo en el modelo, ya que a diferencia de los modelos lineales, este es más flexible. Después de realizar el cálculo:

```
MBE:  1.1597238090192568
```

La diferencia entre el promedio de valores estimados y el promedio de valores reales es de solamente 1.16 unidades, lo cuál no representa un alto sesgo. Lo catalogaría como un sesgo bajo.

Diagnóstico y explicación del grado de varianza

Para la varianza se espera un valor alto, ya que los modelos no lineales como el Árbol de decisión tienden a generar valores con mayor flexibilidad. Para este caso se presentan las siguientes medidas de evaluación de error:

- Esto es lo que se presentó en un primer modelo. Como se puede observar, hay una gran diferencia entre los coeficientes de determinación.

```
--------------MÉTRICAS DE EVALUACIÓN (TRAIN)---------------
RMSE:  0.0
MAE:  0.0
R^2:  1.0

--------------MÉTRICAS DE EVALUACIÓN (TEST)---------------
RMSE:  5.110373807711667
MAE:  4.08340909090909
R^2:  0.33212651521690084
```

- Por otro lado, en el segundo modelo (el primer y segundo modelo serán explicados en la sección de mejoras), se vieron estas medidas:

```
--------------MÉTRICAS DE EVALUACIÓN (TRAIN)----------
RMSE:  3.53625544806577
MAE:  2.728154348177075
R^2:  0.659697979409138

--------------MÉTRICAS DE EVALUACIÓN (TEST)-----------
RMSE:  3.38176563810269
MAE:  2.8349937259709983
R^2:  0.7097438117314442
```

Se puede ver que los valores coinciden en mayor medida y el coeficiente de determinación es mejor.

Ahora analizando lo visto, se puede observar una alta varianza pues al incluir ciertos datos en el primer modelo, el conjunto de entrenamiento es predecido perfectamente (coeficiente igual a 1), esto no es normal, indica un sobreaprendizaje. El modelo cambió completamente con la eliminación de los datos, por lo que indica que jugar con los valores del conjunto de entrenamiento tiene gran impacto a la hora de generar predicciones.

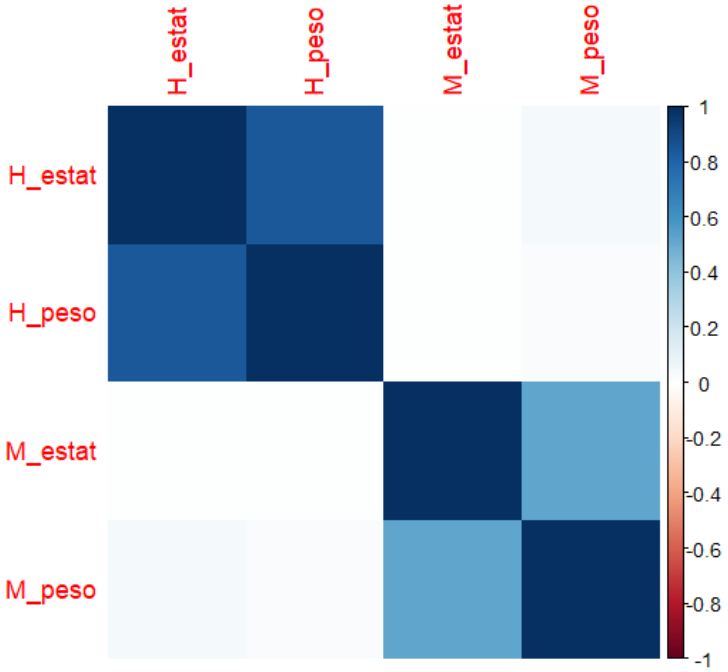
Diagnóstico y explicación el nivel de ajuste del modelo

Según la definición, los árboles de decisión deben tener bajo sesgo y alta varianza, lo que se traduce como un nivel de ajuste alto u overfitting. Esto se confirma, ya que como se vio en el primer modelo elegido, el conjunto de entrenamiento tuvo un coeficiente perfecto, lo cuál no es posible. Y después, las predicciones en su conjunto de pruebas son realmente malas, con un coeficiente de 0.33 (mientras más cercano a 1 mejor).

III.- Mejoras al desempeño

Como se ha mencionado en el análisis, se realizaron dos modelos. El dataset cuenta con cuatro columnas (Peso/estatura en hombres y peso/estatura en mujeres). En el primer modelo se utilizaron todas las columnas para obtener predicciones de los pesos de hombres. Hay que entender que esto no sería muy correcto pues estaría recibiendo influencia de valores que no deberían estar ahí como serían los valores del sexo opuesto. Para contrarrestar el mal primer modelo que se obtuvo, se eligió realizar un pre-procesamiento de los datos.

Para esto se analizó la correlación entre las variables y resultó que evidentemente, la altura y peso de hombres se relaciona.



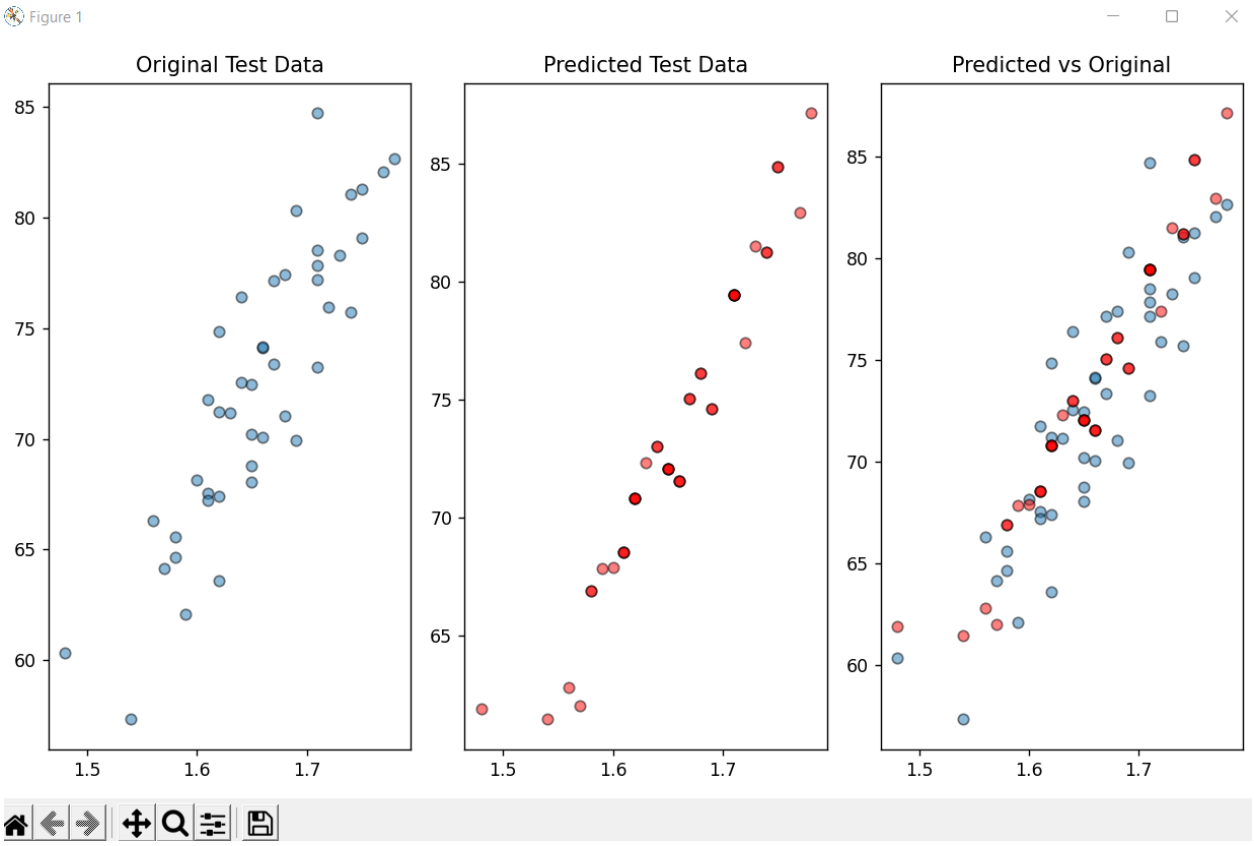
Por esta razón, se quitaron las columnas de los valores del sexo opuesto y el modelo tuvo una mucho mejor ejecución:

```
--------------MÉTRICAS DE EVALUACIÓN (TRAIN)----------
RMSE:  3.53625544806577
MAE:  2.728154348177075
R^2:  0.659697979409138

--------------MÉTRICAS DE EVALUACIÓN (TEST)-----------
RMSE:  3.38176563810269
MAE:  2.8349937259709983
R^2:  0.7097438117314442
```

Una clara mejor en las métricas de evaluación y algo muy importante, no hay mucha diferencia entre lo conseguido por el conjunto de prueba y el de entreno. Además de que el coeficiente de determinación es aceptable.

En gráfica, el mejor modelo elaborado se muestra:



Algunos valores están empalmados, por eso se ven más claros unos que otros. Sin embargo, es evidente que la predicción es positiva.